Intelligent Systems Technologies to Assist in Utilization of Earth Observation Data[†]

Hampapuram Ramapriyan[*a], Gail McConaughy[a], Stephen Morse[b], David Isaac[c]
[a]NASA Goddard Space Flight Center, Greenbelt, MD 20771
[b]SoSA Corporation, Chantilly, VA
[c]Business Performance Systems, Bethesda, MD

ABSTRACT

With the launch of several Earth observing satellites over the last decade, we are now in a "data rich" environment. From NASA's Earth Observing System (EOS) satellites alone, we are accumulating more than 3.5 TB per day of raw data and derived geophysical parameters. The data products are being distributed to a large user community that includes scientific researchers, educators and operational government agencies. Notable progress has been made in the last decade in facilitating access to data. However, to realize the full potential of the growing archives of valuable scientific data, further progress is necessary in the transformation of data into information, and information into knowledge that can be used in particular applications. This paper discusses the concept of an Intelligent Archive in the context of a Knowledge Building system (IA-KBS), with six key capabilities: Virtual Product Generation, Significant Event Detection, Automated Data Quality Assessment, Large-Scale Data Mining, Dynamic Feedback Loop, and Data Discovery and Efficient Requesting. Technologies enabling these capabilities are identified. Many of these technologies are in development today by NSF, NASA and industry sponsorship. These can be taken advantage of for evolving the current generation of data and information systems into the visionary IA-KBS.

Keywords: Intelligent Archives, Data Utilization, Knowledge Building Systems, Data Mining, Knowledge Discovery, Earth Observing System, Remote Sensing, Intelligent Data Understanding, Distributed Systems, Sensor Webs

1. INTRODUCTION

The addition of raw data and derived geophysical parameters from several Earth observing satellites over the last decade to the data held by NASA data centers has created a "data rich" environment for the Earth science research and applications communities. For example, by the end of May 2004, the Distributed Active Archive Centers (DAACs) of NASA's Earth Observing System Data and Information System (EOSDIS) held over 3.4 petabytes of data and derived products containing geophysical parameters. These included over 2100 data product types and 48 million distinct data granules. The data and derived (digital) products were accumulating at the rate of about 3.5 terabytes per day. 2.5 million users accessed DAACs in the year ending in May 2004. The data distribution rate was over 2 terabytes per day.[1]

Today, a "value chain" is in place to take data from satellite observations captured at ground stations from their raw state into intermediate products for scientific research, and to some extent into end products suitable for applications users. There are nearly 80 organizations/systems involved in the value chain including: the EOS Data and Operations System (EDOS); the EOSDIS Core System (ECS); Distributed Active Archive Centers (DAACs); Science Investigator-led Processing Systems (SIPSs); emerging "measurement-based systems" that process measurement sets from multiple missions; Research, Education, and Applications Solutions Network (REASoN) Projects; Federation of Earth Science Information Partners (ESIPs); and the United States Geological Survey's (USGS) and the National Oceanographic and Atmospheric

[†] Copyright 2004 Society of Photo-Optical Instrumentation Engineers. This paper will be published in *Proceedings of SPIE,* **Vol. 5542**, *Earth Observing Systems IX*, and is made available as an electronic preprint with permission of SPIE. One print or electronic copy may be made for personal use only. Systematic or multiple reproduction, distribution to multiple locations via electronic or other means, duplication of any material in this paper for a fee or for commercial purposes, or modification of the content of the paper are prohibited.

[*] This work was performed by the first two authors as part of their duties as U.S. Government employees. It was supported by the Intelligent Systems Project under NASA's Computing, Information and Communications Technology Program. The last two authors worked as subcontractors under Cooperative Agreement NCC5-645 between NASA and George Mason University. The opinions expressed are those of the authors and do not necessarily reflect the official position of NASA.

Administration's (NOAA) long-term archives. This "value chain" is now processing, archiving and distributing data from instruments on board a number of EOS spacecraft, in addition to those from a number of "heritage" missions[2].

As indicated above, the data products are being distributed to a large community of users at a relatively high daily rate. The user community consists of scientific researchers, educators and operational government agencies. Due to technological advances in computational hardware, storage devices, communications networks, and information management software, significant progress has been made over the last two decades in the areas of data archiving and providing the data access for a broad and diverse community of users.

The present state in NASA's Earth science data and information systems has evolved from a rather disparate and disconnected set of data providers through considerable efforts by several organizations. The current state is a very significant improvement over that in the 1980s in terms of the quantities of data managed and ease of access. However, to realize the full potential of the growing archives of valuable scientific data, further progress is necessary in the transformation of data into information, and information into knowledge that can be used in specific applications. Such progress is especially necessary given the projected, even more highly distributed, capabilities - a spectrum that includes sensor webs, distributed processing and archiving environments, and distributed communities of users.

The goals of this paper are: to present some of the issues related to data utilization, to introduce concepts of intelligent archives in the context of knowledge building systems (IA-KBS), to discuss technology needed for an IA-KBS, and to indicate some of the relevant on-going technology developments.

2. DATA UTILIZATION

There are several factors that affect the ability of users to utilize NASA's remotely sensed Earth science data in research as well as operational applications. They include: timeliness, ease of access, understandability, readiness for use, and systems' responsiveness. We will discuss each of the above in the following paragraphs, indicating the present state of systems in meeting the requirements implied by them. This discussion is influenced by, but is not limited to, the present state of EOSDIS and the DAACs.

2.1 Timeliness

The data, information and knowledge must be available to the user at the time when they can be most useful. The timeliness requirements vary considerably depending on whether the data are for research or operational use. Latencies of several days between data acquisition and delivery are generally acceptable for research use. However, operational applications usually require near real-time availability. For instance, operational weather forecasts require data to be available to models within 3 hours of acquisition. However, for development of climate records one can (and must) wait until appropriately calibrated, quality-checked and validated remote sensing data and other ancillary data are available.

Clearly, the near real-time requirements imply significantly more end-to-end system automation, reliability and robustness. Today, given the need to keep up with high rate data flows without building backlogs, most processes, starting from data downlink to archiving of standard products, are automated, even in cases where the user requirements do not call for near real-time availability of data. The data systems are designed to maximize throughput, but not necessarily minimize latency. Operator intervention is usually needed when errors or failures occur, and latency of delivery tends to increase. Additional intelligence in the system can be used to automate recovery from failures, to anticipate and prioritize data requests, and to manage the delivery of near-real-time data through the entire value chain.

2.2 Access

Access requires knowledge of which facility holds the data of interest; the ability to locate all of the data (and/or services) of interest; the ability to obtain no more than what is needed; availability of bandwidth for downloading data electronically or ability to obtain data quickly via media; and appropriate end-to-end capacities for the data flows from the data providers' systems to the users' systems.

Many methods exist today for locating (discovering) data and/or services from a distributed set of providers. Examples of these are: Global Change Master Directory[3], EOS Data Gateway (EDG) [4], Open Source Project for a Network Data Access Protocol (OPeNDAP)[5], Alexandria Digital Library[6], and EOSDIS Clearing House (ECHO)[7]. Access to specialized data products and applications' development for focused user communities have been enabled by NASA through the Federation Experiment involving over 24 Earth Science Information Partners (ESIPs)[8] and, more recently, through Research, Education and Applications Solutions Network (REASoN) Projects. There are a number of efforts underway to take advantage of distributed computing and storage resources that are generally referred to as "Grid Architectures." Examples of these are: the National Science Foundation's National Technology Grid[9], NASA's Information Power Grid[10], the US Department of Energy's DISCOM[11], GriPhyN[12], NEESgrid[13] and the Particle Physics Data Grid[14]. The levels of interoperability, among distributed providers, achieved by these various methods vary, ranging from directory level (providing pointers to sources of data and/or services) to data/service level (providing machine level access to data and services).

Using the goals and characteristics of users in relation to the archive holdings, and awareness of distributed holdings and how they are used can add more intelligence into the archives. Such additional intelligence will help archives respond to users better with more complete, precise and customized responses to data access requests.

2.3 Understandability

The users should clearly understand the salient characteristics of the data, information and knowledge provided to them. They need to know how they were created, what their utility is for the users' particular needs, what their quality attributes (e.g., errors, consistency with previously obtained information, validation status) are and what precautions the users should take in using them.

Some of the steps that have been taken recently in providing Earth science data address the above needs. Examples of these are: documentation accompanying standard products (e.g., Algorithm Theoretical Basis Documents – ATBDs, Guide documents); data quality summaries and metadata; open data policy that includes provision of source code used in generating the products; the ESE's Application Program that collaborates with operational agencies in developing applications of national priority; Education and Outreach programs; and REASoN Projects associated with Education and Applications.

Intelligent archives of the future could improve understandability by assembling and tailoring the wealth of information and knowledge developed and available from various sources and presenting relevant parts of them to the users through appropriate visualization techniques.

2.4 Readiness for use

The data, information and /or knowledge provided to a user must be easy to ingest into the system he/she uses for direct application or manipulation to add value. The closer the item that a user receives is on the "Data-Information-Knowledge" continuum to his/her particular needs, the more useful it will be. It helps if: the items are in the appropriate format(s), reading tools are available, the user can obtain just what he/she needs, and the items are usable with other capabilities (e.g., Geographic Information Systems – GIS) to which the user is accustomed.

Currently, Earth science data typically are delivered in formats that require significant levels of custom programming and manipulation before use. Data are delivered as "granules", or discrete files covering a pre-defined area and time. Most EOS standard products are available in the Hierarchical Data Format (HDF[15]) or its particularization to EOS (HDF EOS[16]), which is very flexible but also fairly complicated and not well supported by geographic information systems and common commercial data analysis tools. Some subsetting (and, more generally, data mining) capabilities exist to enable a user to obtain just what he/she needs from the archived data. However, providing these capabilities for all types of data has been difficult due to the variety of datasets. Another important value-added capability is registration or fusion of data from multiple independent (both temporal and spectral) sources.

Future archives could significantly improve readiness for use by delivering data in a form and format better suited to its intended use, and by providing services that support requests further up the Data-Information-Knowledge continuum.

2.5 Responsiveness

Systems need to be responsive to users' feedback. The word "systems" here refers to hardware, software and people involved in providing data and services. Responsiveness implies making appropriate changes in a timely manner to accommodate users' requests. Examples of required changes are: redirection of sensors to acquire data different from those planned, reprocessing using modified algorithms, implementing a new algorithm or model for computing new products, computations on demand, and implementation of a new client for searching and accessing data.

At present, there are user feedback mechanisms to influence directions of system development and operations. Experience has shown that smaller, focused systems can be more responsive than large and broadly applicable systems – as evidenced by the move to use SIPSs for EOS standard product generation. The development of more heterogeneous and distributed systems will increase the responsiveness, provided there is an appropriate level of coordination as well as implementation of standards and interfaces where necessary. However, there are no "tight" feedback loops that redirect sensors or computational, storage or communications resources depending on user needs or system changes (e.g., partial failures).

Future intelligent archives can play a central role in improving responsiveness by providing value-added mediation services among data collection, processing, and dissemination activities.

3. INTELLIGENT ARCHIVES AND KNOWLEDGE BUILDING SYSTEMS

Given the present state of systems in supporting utilization of data by users discussed above, and projecting the needs through somewhat futuristic scenarios, we have developed a conceptual architecture for an intelligent archive in the context of a knowledge building system (IA-KBS)[17]. The key ideas here are:

- The transformations to go from observations to data to information to knowledge (O_D_I_K) require a feedback loop rather than today's more traditional stovepipe analysis
- Distributed applications are likely to reside in non-traditional platforms in the future (e.g. parts of analyses may happen on-board spacecraft at some future point)
- There is a clear need for more effective distributed infrastructures.

Note that an IA here is assumed to be a distributed set of collaborating systems storing various types and levels of data, information and knowledge. With a KBS supported by an IA, we envision a process in which low-level entering observations (the raw collection output from a sensor) are transformed, in a series of value-added processing steps, first into data, and then into information, and finally into knowledge. As we move across each boundary, the IA serves several supporting functions. First, it serves as the repository, or persistent memory, for the output of a given stage. Thus, when viewed from below, the archive appears as the destination for the results of some output process. However, when viewed above in this cognitive flow, the IA can also appear as the input, or source, for subsequent processing. In this way, by serving as a highly capable buffer, efficient requestor/broker and server of data/ information/ knowledge, the IA can allow the O_D_I_K flow to be spatially and temporally spread. A further level of complexity occurs because, treating the IA as input, a given user or application process may need the IA to provide data (or observations, or information) from a variety of different sources (that is, the upward mapping may be many-to-one). For administrative or system reasons, these inputs may reside at geographically and organizationally distinct locations, and may have different and only partially compatible formats. The IA must then assemble and fuse them (e.g., co-register them), and perhaps also provide further value-added processing before presenting the resulting "virtual product" to the requesting higher-level process.

The efficiency with which these services are provided (e.g., end-to-end latency) can be improved in several ways. For example, an IA can examine its own historical usage patterns, and anticipate changes in levels of request in response to the ebb and flow of various types of geophysical or meteorological events. By predicting these patterns, the IA can assemble and locally cache the inputs for soon-to-be popular products in advance of the actual request(s), and hence reduce communications, processing load and response time in providing service. Individual component systems of the (distributed) IA can cooperate to determine where best to perform processing steps (i.e., moving the algorithm to the data) thereby optimizing the allocation of processing stages to the various participating and co-operating sites.

Another type of service an IA can provide in a KBS-context is event detection and notification. A filter can be assigned to an input stream designed to detect the occurrence of an event (e.g., a forest fire). Notification of the occurrence of the event can then be pushed, in near real-time, to a subscriber along other supporting products to enable rapid and optimized response. This response might include revision to the sensor collection schedule to obtain additional relevant information at the geographical location of interest – a tight sensor-IA-sensor feedback loop.

Another type of service an IA can provide in a KBS-context is to assist in finding archived objects that match the user's research objectives. Searchable metadata containing summaries of pre-computed characteristics is clearly useful, but the IA might also perform specialized content-based searches employing combinations of filters, search criteria, and user-supplied or tailored algorithms. When combined with virtual or fused products from multiple sources, this provides a powerful tool for reducing the users' effort by reducing the volume and increasing the value of the delivered products.

At the top-level of the O_D_I_K chain are data mining algorithms that search for new and perhaps unanticipated relationships and causal connections. In this type of processing, models are at a very high level and may be very broad in spectral, spatial, and temporal characteristics. The discovery and population of such high-level models includes or relies on unsupervised and non-linear techniques that are often highly data- and compute-intensive. By being aware of its role as supplier, partial participant, and (often) recipient of the results of such research approaches, the IA can assist in expediting the extraction of useful knowledge and easing many of the low-level, routine, and data-intensive aspects of this type of research.

4. IA-KBS CAPABILITIES AND TECHNOLOGY NEEDS

There are six key capabilities required in an IA as discussed above. These are: Virtual Product Generation, Significant Event Detection, Automated Data Quality Assessment, Large-Scale Data Mining, Dynamic Feedback Loop, and Data Discovery and Efficient Requesting. In this section, we will discuss technologies needed to enable the vision of an IA-KBS with particular attention to the above six capabilities.

4.1 Virtual product generation (VPG)

An IA does not need to produce and archive all of the products that will be requested of it *in advance*[18]. There are some computational advantages to putting production in the main processing ingest flow. However, the typically, only a small percentage of the products generated are ever actually requested. The alternative is to generate such products "on demand," paying the price of additional latency, but saving computational cycles in the ingest process, and saving the storage costs for these products once produced. If demand can be anticipated (say, via predictive models of usage patterns based on examination of logs), the best of both worlds results. To the user it will appear as if the products were precomputed, but from the archive's point of view, the precomputation is performed only on a relatively small percentage of the data holdings.

Another aspect of VPG is the ability to assemble, transparently, inputs to the production algorithm from a variety of sources and locations. This requires: a global registry (so that potential sources of data are known); configuration management (so that the version numbers of production algorithms are known, and can be kept consistent); and interface standards and supporting middleware (so that production algorithms receive the data in an acceptable and consistent format).

Further, a VPG capability should optimize the assignment of processing resources – both by minimizing the need for data communications, and by taking account of current resource utilization and availability. This means building and using a global predictive model of the interconnected network of storage and processing resources, and keeping the model current via frequent state updates (where "frequent" might mean on the order of 10s of seconds).

The following is a list of enabling and/or closely associated technologies that would be needed to support VPG:

- Data mining (e.g., of usage logs to find predictive usage models)
- Event detection (e.g., to trigger anticipatory production)
- Data fusion and registration

- Computation scheduling and optimization
- Optimized data staging and buffering (e.g., communications)
- Data interface standards, middleware, and distributed metadata
- Distributed system architectures
- Configuration management (e.g., software algorithm version registry and control)

4.2 Significant event detection (SED)

It would be useful if the IA-KBS has the ability to detect significant events that could trigger further actions for the benefit of the users. Suitably designed matched filters or pattern recognition algorithms could be applied to an input data stream (or streams) to detect automatically the occurrence of an event. Some examples of events are: hurricanes, wild fires, volcanic eruptions, failure of some of the sensors in a sensor web, and system failures at one or more of the distributed IA sites. The event detection, in turn, could trigger a variety of actions:

- Notification of subscribers, including real-time latency constraints
- Generation and distribution of associated products
- Retasking of owned or co-operating sensor assets
- Reallocation of system resources
- Self-repair in response to breakage, degradation

The construction of matched filters can itself be a major technical challenge. Typically, a filter on a real-time input stream must be very fast and very accurate, implying very high quality training data and supervised algorithms (e.g., neural nets (NN), Support Vector Machines (SVM)). In order to construct such algorithms however, a prior step is required: the identification and assembly of the training data. This process occurs well in advance while the filter is being developed, and may be the result of content-based retrieval and data mining. In other words, computationally expensive and large-latency data mining can provide the high quality training data input required to construct high-performance, low-latency NNs, SVMs, and other types of real-time filters. Thus, the development of an event-detection capability may be a lengthy process involving a variety of other enabling capabilities and technologies.

The following is a list of a variety of enabling and/or closely associated technologies and capabilities needed to support SED:

- Model building and parameter estimation
- Pattern matching and matched filters (NN, SVM, rule-based, etc.)
- Data mining and content-based retrieval (for extraction of training data)
- System architecture (to facilitate low-latency notification and product generation, scalability)
- State estimation
- Health and status monitoring and instrumentation

4.3 Automated data quality assessment (ADQA)

An IA can take responsibility (to a greater or lesser extent) for monitoring and perhaps correcting the quality of the products it delivers to a requestor or a consuming process[19]. This includes both the algorithms and sources (and their algorithms) used to generate the product (that is, a sophisticated kind of product ancestry and configuration control) as well as internal inspection of the products to ensure that they meet a variety of user-specified characteristics (e.g., cloud-free, dynamic range, sampling resolution, etc.). Other sources of error (e.g., bit errors, compression/decompression lossiness, mistakes in indexes or metadata) can also be detected prior to delivery and, in some cases, corrected or ameliorated (e.g., through the insertion of interpolation or other data modeling approaches).

The following list captures a variety of enabling and/or closely associated technologies and capabilities needed to support ADQA:

- Event detection (the event here could be an error or values exceeding a threshold)
- Health and status monitoring and instrumentation
- System architecture (to support real-time response to error detection and contingent production flows)
- Configuration management

- Data mining (to generate good training sets for error detection algorithms)

4.4 Large-scale data mining (LSDM)

We define LSDM to be a process that finds higher-level emergent causal relationships at a modeling level above the level (in the O_D_I_K hierarchy) at which the inputs exist. Typically, DM sits at the top of the chain – taking information as input, and producing knowledge. Two important examples of this type of analysis are: retrospective studies covering large temporal and geographic extent; and precursor detection, where indicators of significant events are identified through analysis of historical data. Note that, once good precursors have been identified via this scientific LSDM process, computationally efficient filters on real-time data streams can be constructed.

This illustrates that the output of a successful LSDM process can be a set of data, and an associated model, which can serve as the basis for supervised classifiers or event detection algorithms. Often, the model emerging from this process is executable (e.g., decision trees, or rule induction). Once knowledge about a process has been gained, that knowledge can be applied to both new and residual data streams to surface (or predict) otherwise unknown instances of the phenomenon under consideration. The knowledge derived from successful LSDM may have other value or utility: pure science (discovering or confirming previously unverified correlations or relationships – e.g., in global climatology); efficiency (discovering that only some inputs contribute significantly to an output); and instrument or spacecraft health and safety (detection or prediction of anomalies).

There are a number of enabling technologies and algorithmic approaches to LSDM. Further, as we have already noted, one of the goals of an IA is to reduce the burden on the user in extracting narrowly circumscribed archive holdings for which the LSDM is most likely to have success. The following list is representative of the learning techniques typical of LSDM:

- Statistical correlations and entropy minimization
- Principal component analysis
- Independent component analysis
- Clustering
- Automatic rule induction, decision trees, and generalized rule induction
- Bayesian nets and uncertain reasoning
- Supervised techniques (e.g., NNs, SVMs, matched filters)

4.5 Dynamic feedback loop (DFL)

An IA can be modeled[20] as one component in a complete system that includes: sensor tasking, sensors and collection, ground station early level product generation, archiving and higher-level production, real-time or near real-time applications, and user request satisfaction and associated production. There are two low-latency feedback loops of interest. One is the loop that notifies applications of a detected event, which in turn leads to sensor retasking and the triggering of an associated set of events such as data product generation, notification of the user and providing the user with the information he/she needs. This tight, low-latency feedback loop is characteristic of time critical scenarios (for example, the fire detection)

The other type of feedback loop concerns dynamic loading and production optimization, possibly using predictive models to maximize throughput and minimize the latency of delivered user-requested products. This second type of feedback loop requires constructing, populating, and updating system-level models of loading patterns from all sources, and optimizing production schedules over suitably selected sliding time windows taking into account stochastic uncertainties in loading, resource availability, production latency, etc.

The first type of feedback loop is needed *rapid retasking*, and, the second type, for *system resource optimization*.

The following list captures a variety of enabling and/or closely associated technologies and capabilities for DFL:

- Model-predictive control
- Dynamic scheduling and optimization
- Event detection and prediction

- State estimation
- Distributed system architectures
- Non-linear optimized control

4.6 Data discovery and efficient requesting (DDER)

Intelligent archives of the future should be able to detect the presence of newly available data anywhere in the world[21], determine the usefulness of the data, learn how to access them and ultimately provide the data to users or applications in a usable form. This involves: an ongoing Google-like search process that keeps constantly and persistently aware of potential data sources and their changing status; and capabilities to retrieve and reformat the data transparently to meet the users' data interface requirements or preferences. In this way, the IA becomes an interface not only to its own holdings, but also to the broad array data sources of interest across the accessible web.

There are aspects of this capability that overlap VPG (see section 4.1 above), since, as in the case of virtual products, the user need not be explicitly aware of the underlying sources for the delivered product. In the discussion of VPG, however, the emphasis was on delaying creation of "standardized" products until (or, perhaps, shortly prior to) the actual request, and optimizing the associated production process. Here, the concern is not so much on production as on transparent access, retrieval, and reformatting. Further, an Efficient Requestor may also contain tasking models of sensor systems, and could perhaps broker highly efficient sensor schedules opportunistically so as to optimize system-level value delivered.

The following list captures several capabilities and technologies that enable DDER or are closely associated with it:

- Agent-based search
- Middleware and data interface standardization
- Distributed indexes and registries
- Virtual product generation (and its technologies)
- User profiling (e.g., to determine data format delivery preferences)
- State estimation
- Automated, optimized planning and scheduling

5. RELEVANT TECHNOLOGY DEVELOPMENTS

There are several technology developments occurring today that will help realize the vision of the IA-KBS in the future. The developments are at varying levels of maturity and the development of an IA-KBS will be a gradual, evolutionary process. Some of the on-going research activities relevant to IA-KBS are briefly discussed in this section.

Significant investment is being made by the National Science Foundation (NSF) and industry research community in the area of "emerging distributed architectures". Many of the elements of these research efforts are related to "grid" architectures and are being applied by NSF to discipline-based science community data and resource sharing (e.g. TeraGrid[22], GEON[23], and OptiPuter[24] (California Institute for Telecommunications and Information Technology's effort to re-optimize the grid stack of software abstractions). In addition, industry is announcing grid-based products, for example IBM's utility computing[25] and Oracle's grid-based databases[26]. Other community efforts supporting distributed architectures include web service based concepts (e.g. GSFC's ECHO[4]), and address data interoperability (e.g. University of Rhode Island's OPeNDAP[5]).

NASA's Intelligent Systems Project[27] within the Computing, Information and Communication Technology (CICT) Program has been supporting development of some of the technologies needed. Examples of these that are particularly relevant to IA-KBS from the point of view of Earth science are given below:

Fern and Brodley[28, 29] have investigated "Machine Learning and Data Mining for Intelligent Data Understanding of High Dimensional Earth Science Data" using techniques including unsupervised clustering, random projection onto lower dimensional subspaces, ensemble method clustering, and bipartite graph partitioning to solve the cluster ensemble problem.

Kumar et al[30, 31] have researched into the "Discovery of Changes from the Global Carbon Cycle and Climate System using Data Mining" with the goal of demonstrating use of clustering techniques on time series data to recover climate indices. They use shared nearest neighbor clustering, area weighted temporal correlation metrics and standard techniques in statistical analysis (e.g., normalization, reduction of trends, accounting for temporal lag)

Teng's research[32, 33], "Polishing: Enhancing Data Quality by Repairing Imperfections" is aimed at using a classifier and training data to detect anomalies in the feature data, selectively correcting the anomalies, and retraining the classifier with the altered data to obtain improved classifier performance.

Kargupta[34, 35] has investigated "Distributed Data Mining for Large NASA Databases". He uses techniques such as Distributed Principal Components Analysis (DPCA), Distributed Bayesian Nets (DBN), distributed clustering, and distributed randomized inner product (DRIP) to handle large datasets located at distributed sites.

Smelyanskiy's research[36] into "Statistical Modeling, Forecasting and Control of Large Deviation Events" uses Bayesian inference (that is, the construction of an appropriate likelihood function) to infer a parameterized model of a non-linear stochastic dynamical system from observed time series data.

Nemani and Golden[37] have investigated "Ecosystem Forecasting". With their research, given a user request for a "product," generally conceived of as a graphical map overlay, the system identifies the inputs (or input alternatives) and the intermediate processing required to generate the product, optimizes the production run, and delivers the requested data in the desired format. The system also has the ability to learn how to generate new products by observing processing sequences supplied by users.

LeMoigne's research[38, 39] into "Image Registration and Fusion for Future Formation Flying Systems" has the goals of achieving sub-pixel accuracies for image-to-image registration; developing a framework for performing trade-offs among alternative registration component technologies; and reduction in dimensionality of hyper-spectral data.

6. CONCLUSION

The addition of raw data and derived geophysical parameters from several Earth observing satellites over the last decade to the data held by NASA data centers has created a "data rich" environment for the Earth science research and applications communities. Due to technological advances in computational hardware, storage devices, communications networks, and information management software, significant progress has been made over the last two decades in the areas of data archiving and providing the data access for a broad and diverse community of users. Today, a "value chain" is in place to take data from satellite observations captured at ground stations from their raw state into intermediate products for scientific research, and to some extent into end products suitable for applications users. However, to realize the full potential of the growing archives of valuable scientific data, further progress is necessary in the transformation of data into information, and information into knowledge that can be used in specific applications. Such progress is especially necessary given the projected, even more highly distributed, capabilities - all the way from sensor webs to distributed processing and archiving environments to distributed communities of users.

There are several factors that affect the ability of users to utilize NASA's remotely sensed Earth science data in research as well as operational applications. They include: timeliness, ease of access, understandability, readiness for use, and systems' responsiveness. Knowledge Building Systems (KBS), and specifically, Intelligent Archives in the context of KBS (IA-KBS) can help advance the state-of-the-art in utilization of data significantly. An IA-KBS is conceived here as a distributed set of collaborating systems storing various types and levels of data, information and knowledge. With a KBS supported by an IA, we envision a process in which low-level entering observations (the raw observational output from a sensor) are transformed, in a series of value-added processing steps, first into data, and then into information, and finally into knowledge. Six key capabilities of an IA-KBS are: Virtual Product Generation, Significant Event Detection, Automated Data Quality Assessment, Large-Scale Data Mining, Dynamic Feedback Loop, and Data Discovery and Efficient Requesting. We have identified a number of technologies needed to enable these key capabilities.

There are several technology developments occurring today that will help realize the vision of the IA-KBS in the future. The developments are at varying levels of maturity and the development of an IA-KBS will be a gradual, evolutionary process. Relevant developments include those resulting from investments from NSF and industry in the area of emerging distributed architectures, from NASA's Applied Information Systems Technology (AIST) program within the Earth Science Technology Office (ESTO), and from NASA's Intelligent Systems Project within the Computing, Information and Communication Technology (CICT) Program. These technologies need to be incorporated into a testbed environment to move into next levels of maturity. Especially important to consider is the scaling of the techniques and algorithms that have been tested on small sample datasets to larger "real-world" problems[40].

ACKNOWLEDGMENT

This paper was a result of the work by the IA-KBS study team consisting of the authors and the following individuals: S. Kempler, C. Lynnes, and K. McDonald of NASA Goddard Space Flight Center; and L. Di, W. Yang and P. Zhao of George Mason University.

REFERENCES

1. Y-C. Lu, "EDGRS (ESDIS Data Gathering and Reporting System)", http://fiacre:8000/, NASA Goddard Space Flight Center, Dynamic Website.
2. H. K. Ramapriyan, "NASA's Earth Science Data Systems – Past, Present and Future", *Proceedings of the IGARSS 2003 Conference*, Toulouse, France, pp I:637-639, 2003.
3. L. M. Olsen, "Discovering and using global datasets", in *Global Environmental Databases: Present Situation; Future Directions,* R. Tateishi and D. Hastings editors, International Society for Photogrammetry and Remote Sensing (ISPRS), Geocarto International Center, Hong Kong, pp 220-233, 2000.
4. R. Pfister, "The Information Management System of NASA's EOS Data and Information system", *Proceedings of the American Meteorological Society (AMS) 17th International Conference on Interactive Information and Processing Systems (IIPS) for Meteorology, Oceanography, and Hydrology*, Albuquerque, New Mexico, pp 161-163, 2001.
5. P. Cornillon, "OPeNDAP", http://opendap.org, OPeNDAP, Inc., Dynamic Website.
6. J. Frew, M. Freeston, L. Hill, G. Janée, M. Larsgaard, Q. Zheng, "Generic Query Metadata for Geospatial Digital Libraries," *Proceedings of the Third IEEE META-DATA Conference*, 1999.
7. R. Pfister and K. Weichman, "New Paradigm For Search and Order in EOSDIS", *Proceedings of the IEEE 2000 International Geoscience and Remote Sensing Symposium*, Honolulu, Hawaii, 2000.
8. D. Jones and D. Hardin, "Federation of Earth Science Information Partners", http://www.esipfed.org/, Dynamic Website.
9. Smarr, L., "The Emerging National Technology Grid: Using High-End Computational Science to Predict the Broad-Based Future", http://www.jacobsschool.ucsd.edu/~lsmarr/talks/San%20Diego/index.htm, December 1998.
10. W. E. Johnston, L. A. Tanner, W. J. Feiereisen, W. Thigpen, "Information Power Grid: Distributed High-Performance Computing and Large-Scale Data Management for Science and Engineering", http://www.globus.org/retreat00/presentations/o_IPG.ProtoTestb.pdf, July 2000.
11. DISCOM, "Distance and Distributed Computing and Communication", http://www.cs.sandia.gov/discom/main.html, 2002.
12. Grid Physics Network, http://www.griphyn.org/projinfo/index.php, Dynamic Website.
13. NEESgrid, http://www.neesgrid.org/, Dynamic Website.
14. PPDG, "Particle Physics Data Grid", http://www.ppdg.net/, Dynamic Website.
15. NCSA, "The NCSA HDF Home Page: Information, Support, and Software from the Hierarchical Data Format (HDF) Group of NCSA", Dynamic Website.
16. R. Ullman and L. J. Tyhala, "HDF-EOS Tools and Information Center", http://hdfeos.gsfc.nasa.gov/, Dynamic Website.
17. H. K. Ramapriyan, G. McConaughy, C. Lynnes, R. Harberts, L. Roelofs, S. Kempler, and K. McDonald, "Intelligent Archive Concepts for the Future", *Proceedings of the ISPRS/Future Intelligent Earth Observing Systems Conference*, Denver, CO, 2002.
18. M. Clausen, M. and C. Lynnes, "Virtual Data Products in an Intelligent Archive", http://daac.gsfc.nasa.gov/IDA/presentations.shtml, 2003.
19. D. Isaac, and C. Lynnes, "Automated Data Quality Assessment in the Intelligent Archive", http://daac.gsfc.nasa.gov/IDA/presentations.shtml, 2003.

20. H. Morse, and D. Isaacs, “Optimizing Performance in Intelligent Archives”, http://daac.gsfc.nasa.gov/IDA/presentations.shtml, 2003.
21. C. Lynnes, “Automated Data Discovery and Usage”, http://daac.gsfc.nasa.gov/IDA/presentations.shtml, 2003.
22. NCSA, “About Teragrid”, http://archive.ncsa.uiuc.edu/About/TeraGrid/, 2002.
23. P. Tooby, “GEON Overview: Cyberinfrastructure for Geosciences”, http://www.geongrid.org/, 2003.
24. OptIPuter, “A Powerful Distributed Cyberinfrastructure to Support Data-Intensive Scientific Research and Collaboration”, http://www.optiputer.net/, 2003.
25. P. Korzeniowski, “IBM’s Future Strategy: Grid Computing Everywhere”, http://www.ecommercetimes.com/story/31269.html, 2003.
26. M. Miley “Bringing Computing Power to the Grid”, http://otn.oracle.com/oramag/oracle/03-sep/o53grid.html, 2003.
27. NASA, “Computing, Information and Communications Technology: Intelligent Systems Program”, http://is.arc.nasa.gov/, Dynamic Website.
28. X. Fern, C. Brodley, “Solving cluster ensemble problems by bipartite graph partitioning,” *Proceedings of the twenty-first International Conference on Machine Learning*, Banff, Canada, 2004.
29. X. Fern, C. Brodley, “Random projection for high dimensional data clustering: A cluster ensemble approach,” *Proceedings of the Twentieth International Conference on Machine Learning*, Washington DC, 2003.
30. M. Steinbach, P. Tan, V. Kumar, S. Klooster, C. Potter, “Discovery of Climate Indices using Clustering”, *KDD 2003*, Washington, DC, USA, 2003.
31. C. Potter, P. Tan, M. Steinbach, S. Klooster, V. Kumar, R. Myneni, V. Genovese, “Major disturbance events in terrestrial ecosystems detected using global satellite data sets”, *Global Change Biology*, **9,** pp 1005-1021, 2003.
32. C. M. Teng, “Correcting noisy data,” *Technical Report, UNSW* (Sydney, Australia), 1998.
33. C. M, Teng, “Polishing Blemishes: Issues in Data Correction,” *IEEE Intelligent Systems,***19:2**, pp 34 – 39, March/April 2004.
34. R. Chen, K. Sivakumar, and H. Kargupta, “Collective Mining of Bayesian Networks from Heterogeneous Data”, *Knowledge and Information Systems Journal*, **6**, pp 164-187, 2001.
35. H. Kargupta, W. Huang, S. Krishnamoorthy, and E. Johnson, “Distributed Clustering Using Principal Component Analysis”, *Knowledge and Information Systems Journal*, **5,** pp 422-448, 2000.
36. V. Smelyanskiy, Timucin, D. A., Bandrivskyy, A., Luchinsky, D. G., “Model reconstruction of nonlinear dynamical systems driven by noise,” *Archiv:Physics*/0310062, October 2003.
37. K. Golden, W. Pang. R. Nemani, P. Votava, “Automating the Processing of Earth Observation Data”, *Proceeding of seventh International Symposium on Artificial Intelligence, Robotics and Automation for Space, 2003.*
38. S. Kaewpijit, J. LeMoigne, T. El-Ghazawi, “Automatic Reduction of Hyperspectral Imagery Using Wavelet Spectral Analysis”, *IEEE Transactions on Geoscience and Remote Sensing*, **41,** pp 863-871, 2003.
39. J. LeMoigne, N. Netanyahu, J. Masek, “Geo-Registration of Landsat Data via Robust Matching of Multi-Resolution Features,” *IEEE Transactions on Geoscience and Remote Sensing*, [in press], 2004.
40. G. McConaughy, and K. McDonald, “Moving from Data and Information Systems to Knowledge Building Systems: Issues of Scale and Other Research Challenges”, http://daac.gsfc.nasa.gov/IDA/presentations.shtml, 2003.